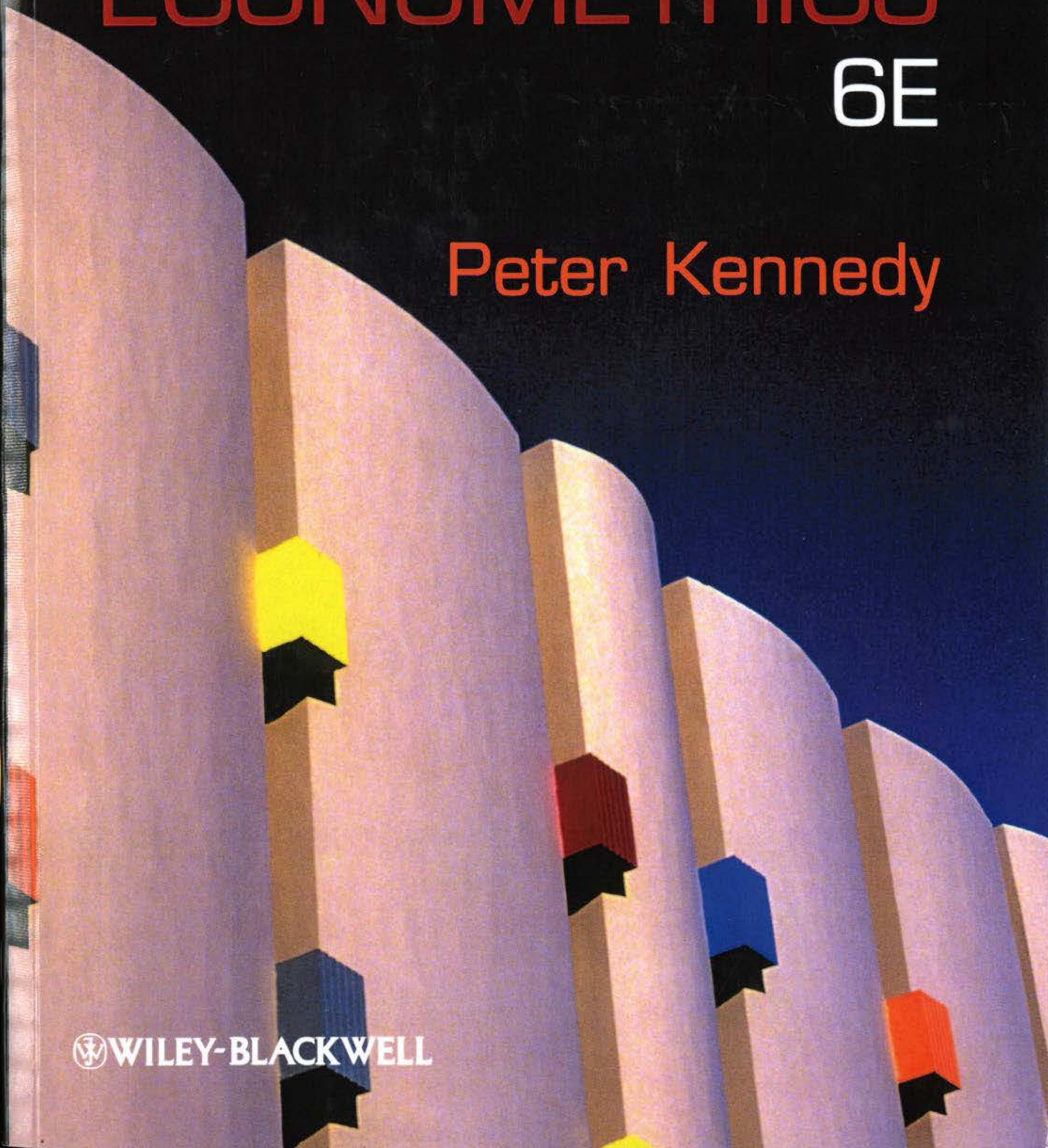
Exhibit 13

A Guide to ECONOMETRICS

6E

Peter Kennedy



WILEY-BLACKWELL

BLACKWELL PUBLISHING
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford OX4 2DQ, UK
550 Swanston Street, Carlton, Victoria 3053, Australia

First edition published 1979 by Martin Robertson and Company Ltd
Second and Third editions published by Blackwell Publishers Ltd
Fourth edition published 1998
Fifth edition published 2003
Sixth edition published 2008 by Blackwell Publishing Ltd

5 2011

Library of Congress Cataloging-in-Publication Data

Kennedy, Peter, 1943–
A guide to econometrics / Peter Kennedy. — 6th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4051-8258-4 (hardcover : alk. paper) — ISBN 978-1-4051-8257-7 (pbk. : alk. paper)
1. Econometrics. I. Title.

HB139.K45 2008
330.01'5195—dc22
2007039113

A catalogue record for this title is available from the British Library.

Set in 10.5/12.5 pt Times
by Newgen Imaging Systems (P) Ltd, Chennai, India

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

For further information on
Blackwell Publishing, visit our website:
www.blackwellpublishing.com

MIX
Paper from responsible sources
FSC www.fsc.org FSC® C013604

Chapter 9

Violating Assumption Four: Instrumental Variable Estimation

9.1 Introduction

The fourth assumption of the classical linear regression (CLR) model specifies that the observations on the explanatory variables can be considered fixed in (conceptual) repeated samples. In many economic contexts the explanatory variables are themselves random/stochastic variables and thus could not possibly have the same values in repeated samples. A classic example is a simultaneous equation system with supply and demand curves. To estimate the demand curve we would regress quantity on price, among other variables. When we draw new error terms for the supply and demand equations to create a repeated sample, the intersection of the supply and demand curves changes and so the price changes: price is stochastic, it cannot remain fixed in repeated samples.

This assumption of fixed regressors is made mainly for mathematical convenience; if the regressors can be considered to be fixed in repeated samples, the desirable properties of the ordinary least squares (OLS) estimator can be derived quite straightforwardly. The role of this assumption in derivations of OLS estimator properties is to make the regressors and errors independent of one another. If this assumption is weakened to allow the explanatory variables to be stochastic but to be distributed independently of the error term, all the desirable properties of the OLS estimator are maintained; their algebraic derivation is more complicated, however, and their interpretation in some instances must be changed (for example, in this circumstance β^{OLS} is not, strictly speaking, a linear estimator). Even the maximum likelihood property of β^{OLS} is maintained if the disturbances are distributed normally and the distribution of the regressors does not involve the unknown parameters β and σ^2.

This fourth assumption can be further weakened at the expense of the small-sample properties of β^{OLS}. If the regressors are *contemporaneously uncorrelated* with the disturbance vector, the OLS estimator is biased but retains its desirable asymptotic properties. Contemporaneous uncorrelation in this context means that the *n*th observation on

all regressors must be uncorrelated with the nth disturbance term, but it is allowed to be correlated with the disturbance terms associated with other observations. Suppose, for example, that a lagged value of the dependent variable, lagged y, appears as one of the explanatory variables. When we draw a new vector of error terms to create a repeated sample, all the dependent variable values, including lagged y, change because the error is a part of the equation determining the dependent variable. So the value of lagged y, one of the explanatory variables, is stochastic and cannot be considered as fixed in repeated samples. But in this example, although lagged y is correlated with the error in its own time period, period $t-1$, it is not correlated with the error in the following period, period t. The error in the equation being estimated is the error for period t, so no contemporaneous correlation exists between lagged y and the regression error. OLS will be biased, but consistent. In this case no alternative estimators are available with superior small-sample properties, so the OLS estimator is retained on the basis of its desirable asymptotic properties. Henceforth the "contemporaneous" qualification is dropped for expositional ease, so that the terminology "regressor correlated with the error" means contemporaneous correlation.

If the regressors are correlated with the error term, the OLS estimator is biased even asymptotically. (And this bias in general spills over to the estimates of all the slope coefficients, not just the slope of the regressor creating the problem!) The bias happens because the OLS procedure, in assigning "credit" to regressors for explaining variation in the dependent variable, assigns, in error, some of the disturbance-generated variation of the dependent variable to the regressor with which that disturbance is correlated. Consider as an example the case in which the correlation between the regressor and the disturbance is positive. When the disturbance is higher the dependent variable is higher, and owing to the correlation between the disturbance and the regressor, the regressor is likely to be higher, implying that too much credit for making the dependent variable higher is likely to be assigned to the regressor. This is illustrated in Figure 9.1. If the error term and the independent variable are positively correlated, negative values

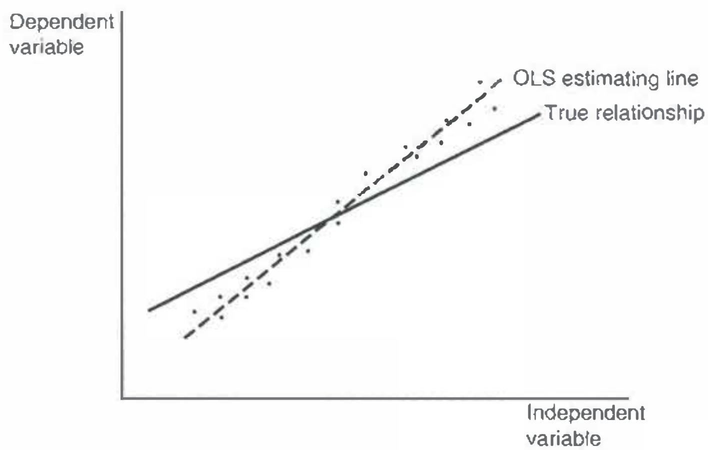


Figure 9.1 Positive contemporaneous correlation.

but it is allowed to be
vations. Suppose, for
appears as one of the
is to create a repeated
ange because the error
o the value of lagged
considered as fixed in
related with the error
error in the following
e error for period t, so
regression error. OLS
tors are available with
ined on the basis of its
eous" qualification is
or correlated with the

estimator is biased even
imates of all the slope
lem!) The bias happens
s for explaining varia-
disturbance-generated
that disturbance is cor-
n between the regressor
r the dependent variable
ce and the regressor, the
or making the dependent
illustrated in Figure 9.1.
rrelated, negative values

estimating line
True relationship

dependent
variable

of the disturbance will tend to correspond to low values of the independent variable and positive values of the disturbance will tend to correspond to high values of the independent variable, creating data patterns similar to that shown in the diagram. The OLS estimating line clearly overestimates the slope of the true relationship. (This result of overestimation with positive correlation between the disturbance and regressor does not necessarily hold when there is more than one explanatory variable, however; the pattern of bias in the multivariate case is complicated.) Note that the estimating line provides a much better fit to the sample data than does the true relationship; this causes the variance of the error term to be underestimated.

When there is correlation between a regressor and the error term, that regressor is said to be *endogenous*; when no such correlation exists the regressor is said to be *exogenous*. Endogeneity gives rise to estimates biased even asymptotically, making economists very unhappy. Indeed, this is one of the features of economic data that distinguishes econometrics from other branches of statistics. The heart of the matter is that the data with which econometricians work seldom come from experiments designed to ensure that errors and explanatory variables are uncorrelated. Here are some examples of how this endogeneity problem can arise.

Measurement error in explanatory variables. Suppose $y = \alpha + \beta x + \varepsilon$ but that measured x, x_m, is $x + u$ where u is a random error. Add and subtract βu to find that the relationship between y and x_m, the explanatory variable used in the regression, is $y = \alpha + \beta x_m + (\varepsilon - \beta u)$. When a repeated sample is taken it must involve new values for measurement errors u in the data, as well as new values for the traditional error term ε. But clearly u affects both x_m and the composite error $(\varepsilon - \beta u)$; in this regression there is correlation between the explanatory variable x_m and the error term $(\varepsilon - \beta u)$. The general topic of measurement errors is discussed in chapter 10.

Autoregression with autocorrelated errors. Suppose the lagged value of the dependent variable, lagged y, is a regressor. When new errors are drawn for a repeated sample all values of the dependent variable change, including lagged y, so this regressor is stochastic. As noted earlier, lagged y is not contemporaneously correlated with the regression error. If the errors are autocorrelated, however, then this period's error is correlated with last period's error. But last period's error is a direct determinant of lagged y: this creates correlation between lagged y and this period's error. An obvious lesson here is that whenever lagged y appears as a regressor we should test for autocorrelated errors! Autoregression is discussed further in chapter 10.

Simultaneity. Suppose we are estimating a demand curve so that one of the explanatory variables is price. If the error term in this equation bumps up it shifts the demand curve and so through its simultaneity/intersection with the supply curve changes the price. This causes correlation between the demand curve errors and the explanatory variable price. In general, all endogenous variables in a system of simultaneous equations are correlated with all of the errors in that system. Simultaneity, sometimes referred to as *reverse causation*, is very common in econometric work.

Changes in policing, for example, could cause changes in crime rates, but changes in crime rates prompt changes in policing. Suppose we regress crime on policing. When the error term in this regression bumps up it directly increases crime. This increase in crime increases policing through the reverse causation (the simultaneity). This means that the error term is correlated with policing, so when we regress crime on policing we get biased estimates. In effect, when we regress crime on policing, some of the reverse influence of crime affecting policing gets into the coefficient estimate, creating simultaneity bias. Simultaneity is discussed at length in chapter 11.

Omitted explanatory variable. Whenever an explanatory variable has been omitted its influence is encompassed by the error term. But often the omitted explanatory variable is correlated with included explanatory variables. This makes these included explanatory variables correlated with the error term. Suppose, for example, that we are regressing wage on years of education but do not have an ability variable to include in the regression. People with higher ability will earn more than others with the same years of education, so they will tend to have high positive error terms; but because they have higher ability they will find it worthwhile to take more years of education. This creates correlation between the error term and the explanatory variable years of education. This is another way of viewing the omitted variable specification error discussed earlier in chapter 6.

Sample selection. Often people appear in a sample because they have chosen some option that causes them to be in the sample. Further, often this choice is determined by characteristics of these people that are unmeasured. Suppose you are investigating the influence of class size on student learning. Some parents may have gone to a lot of trouble to ensure that their child is in a small class; parents who take this trouble probably are such that they work hard with their child at home to enhance their child's learning, and thereby create for that child, other measured things being equal, a positive error term. A consequence of this sample selection phenomenon is that small classes are more likely to be populated by children with positive errors, creating (negative) correlation between class size and the error in the learning equation. This problem is sometimes referred to as *unobserved heterogeneity*; the observations in the sample are heterogeneous in unobserved ways that create bias. This is related to the omitted explanatory variable category above; if we could measure the causes of the heterogeneity we could include these measures as extra explanatory variables and so eliminate the bias. This sample selection problem is also addressed later in chapter 16 when discussing limited dependent variables.

The bottom line here is that correlation between explanatory variables and a regression's error term is not an unusual phenomenon in economics, and that this is a serious problem because it introduces bias into the OLS estimator that does not disappear in large samples. Unfortunately, there does not exist an alternative estimator which is unbiased; the best we can do is turn to estimation procedures that are unbiased asymptotically, or defend OLS using the mean square error (MSE) criterion. The purpose of this chapter is to exposit the instrumental variable (IV) estimator, the most common

estimator employed as an alternative to OLS in this context. Why is a whole chapter devoted to this estimator? There are several reasons for this. First, this procedure is one which has a rich history in econometrics, and its frequent use is a major way in which econometrics is distinguished from other branches of statistics. Second, the procedure permeates a large part of econometrics in various ways, so a good understanding of it is of value. And third, there are lots of issues to be dealt with, some of which are quite problematic: How does the technique work? How do we find the instruments it requires? How do we test if these instruments are any good? How do we interpret the results? We begin by describing the IV procedure.

9.2 The IV Estimator

The IV procedure produces a consistent estimator in a situation in which a regressor is correlated with the error, but as noted later, not without cost. To facilitate exposition henceforth regressors that are correlated with the error are referred to as "troublesome" or "endogenous" explanatory variables. To use the IV estimator one must first find an "instrument" for each troublesome regressor. (If there is not at least one unique instrument for each troublesome variable the IV estimation procedure is not *identified*, meaning that it cannot produce meaningful estimates of the unknown parameters. Not being identified is like having more unknowns than equations – the equations can't be solved, in the sense that there is an infinite number of values of the unknowns that satisfy the equations!) This instrument is a new independent variable which must have two characteristics. First, it must be uncorrelated with the error; and second, it must be correlated (preferably highly so) with the regressor for which it is to serve as an instrument. The IV estimator is then found using a formula involving both the original variables and the IVs, as explained in the technical notes. (It is *not* found by replacing the troublesome variable with an instrument and running OLS, as too many students believe!) The general idea behind this estimation procedure is that it takes variation in the explanatory variable that matches up with variation in the instrument (and so is uncorrelated with the error), and uses only this variation to compute the slope estimate. This in effect circumvents the correlation between the error and the troublesome variable, and so avoids the asymptotic bias.

A major drawback to IV estimation is that the variance of the IV estimator is larger than that of the OLS estimator. It is easy to see why this happens. As explained earlier, only a portion of the variation in the troublesome variable (the portion matching up with the instrument) is used to produce the slope estimate; because less information is used, the variance is larger. This portion is larger (and so the variance smaller) the larger is the correlation between the troublesome variable and the instrument; this is why the "preferably highly so" was included earlier. This higher variance, which is sometimes dramatically higher, is the price paid for avoiding the asymptotic bias of OLS; the OLS estimator could well be preferred on the MSE criterion. A second major drawback is that if an instrument is "weak," meaning that its correlation with the troublesome variable is low, as explained later the IV estimates are unreliable, beyond just having a high variance.